Towards Automatic Generation of Multimodal Answers to Medical Questions: A Cognitive Engineering Approach

Charlotte van Hooijdonk, Emiel Krahmer, Alfons Maes
Tilburg University
Tilburg, The Netherlands
{c.m.j.vanhooijdonk | e.j.krahmer | maes}@uvt.nl

Mariët Theune, Wauter Bosma
University of Twente
Enschede, The Netherlands
{m.theune | w.e.bosma}@utwente.nl

Abstract

This paper describes a production experiment carried out to determine which modalities people choose to answer different types of questions. In this experiment participants had to create (multimodal) presentations of answers to general medical questions. The collected answer presentations were coded on types of manipulations (typographic, spatial, graphical), presence of visual media (i.e., photos, graphics, and animations), functions and position of these visual media. The results of a first analysis indicated that participants presented the information in a multimodal way. Moreover, significant differences were found in the information presentation of different answer and question types.

Keywords: Multimodal information presentation, cognitive engineering

1 INTRODUCTION

Much research on question answering (QA) has focussed on answering factoid questions, i.e., questions that have one word or phrase as their answer, such as "Amsterdam" in response to the question "What is the capital of the Netherlands?" Obviously, factoid QA does not really require Natural Language Generation, and the output modality will typically be text. However, there is currently a growing interest in moving beyond factoid questions and purely textual answers, and then output generation becomes an important issue. Questions that arise are: how to determine for a given question, what the best combination of modalities for the answer is? And related to this: what is the proper length of a non-factoid answer? In this paper, we describe ongoing work in the context of a medical QA system within the IMIX / IMOGEN[1] project that addresses exactly these issues.

In the medical domain several question types occur, such as definition questions or procedural questions. These different types of questions require different types of answers. For example the answer to the definition question "What does RSI stand for?" would probably be a textual answer, like "RSI stands for Repetitive Strain Injury". However, the presentation of an answer through text only may not be the best choice for every type of information. In some cases other modalities (e.g., pictures, film clips, etc.) or modality combinations (e.g., text + picture) may be more suitable. For example the answer to the procedural question "How should I organize my workspace in order to prevent RSI?" would probably be more informative if it contained a picture. Moreover, the length of the answer could also play an important role in the answer presentation. For example, the answer to the question "What does RSI stand for?" could be an extended one: "RSI stands for Repetitive Strain Injury. This disorder involves damage to muscles, tendons and nerves caused by overuse or misuse, and affects the hands, wrists, elbows, arms, shoulders, back, or neck". This answer provides the user with relevant background information about the topic of the

[1] For more information about IMOGEN, see http://wwwhome.cs.utwente.nl/~theune/IMOGEN/index.html/.

question. In addition, including additional text in the answer may allow the user to assess the answer's accuracy in order to verify whether it is correct or not (Bosma, 2005). This raises the question which kind of answer presentations (unimodal vs. multimodal) would be best for different types of questions and answers.

Much research has been done in the field of cognitive psychology on the influence of (combinations of) different modalities on the users' understanding, recall and processing efficiency of the presented material (e.g., Carney & Levin 2002, Mayer 2005, Tversky, Morrison & Betrancourt 2002). This research has resulted in several guidelines on how to present (multimodal) information to the user, such as the multimedia principle (i.e., instructions should be presented using both text and pictures, rather than text only) and the spatial contiguity principle (i.e., when presenting a combination of text + pictures, the text should be close to or embedded within the pictures) (Mayer, 2005). However, these guidelines are based on specific types of information used in specific domains in particular descriptions of cause and effect chains which explain how systems work (Mayer 1989, Mayer & Gallini 1990, Mayer & Moreno 2002) and procedural information describing how to acquire a certain skill (Marcus, Cooper & Sweller 1996, Michas & Berry 2000, Schwan & Riempp 2004). These guidelines do not tell us which modalities are most suited for which information types, as each learning domain has its own characteristics (Van Hooijdonk & Krahmer, submitted).

Several researchers have tried to make an overview of the characteristics of modalities, information types, and the matches between them. For example, Bernsen (1997) focussed on the features of modalities in his Modality Theory, i.e., *"given any particular set of information which needs to be exchanged between user and system during task performance in context, identify the input/output modalities, which, from the user's point of view, constitute an optimal solution to the representation and exchange of that information"*. He proposed a taxonomy to define generic unimodalities consisting of various features. Other researchers proposed taxonomies of information types such as dynamic, static, conceptual, concrete, spatial, and temporal in order to select the appropriate modalities (e.g., Heller, Martin, Haneef, Gievska-Krliu 2001, Sutcliffe, 1997).

Other research has been concerned with the so-called media allocation problem: *"How does a producer of a presentation determine which information to allocate to which medium, and how does a perceiver recognize the function of each part as displayed in the presentation and integrate them into a coherent whole?"* (Arens, Hovy & Vossers, 1993). According to Arens et al. (1993) the characteristics of the media used are not the only features that play a role in media allocation. The characteristics of the information to be conveyed, the goals and characteristics of the producer, and the characteristics of the perceiver and the communicative situation are also important. In order to create a multimodal information presentation, modalities should be integrated dynamically based on a communication theory as a whole (e.g., André 2000, Arens et al. 1993, Maybury & Lee 2000, Oviatt et al. 2003).

In short, several research fields have been concerned with the generation of multimodal information presentations resulting in several guidelines, frameworks, and taxonomies. However what is really needed to generate optimal multimodal presentations is gaining knowledge on whether users present information in a multimodal way, and if so, when and how they present this multimodal information. To achieve this goal, we will carry out a series of experiments following the cognitive engineering approach as used by Heiser et al. (2004). In this approach, human users are asked to produce information presentations, which are then rated by other users. Based on the results, cognitive design principles are identified and used to improve the automatic generation of information presentations. In this paper, we present a production experiment in which users' (multimodal) answers to different medical questions were collected. We expected that both question type (definition vs. procedural questions) and answer type (brief or extended) would affect the answer presentation, i.e., some answers would probably consist of text only while others would probably consist of a combination of modalities like text + picture. In a later stage, a preference experiment will be conducted in which other users will rate the answer presentations collected in the production experiment.

This paper is structured as follows. In section 2 the research method is described followed by the results of a first analysis in section 3. We end with a general conclusion in section 4.

2 METHOD

2.1 PARTICIPANTS

One hundred and eleven students of Tilburg University participated for course credits. Of the students, 46 were male and 65 were female. The mean age was 22 years (Std = 2,10). All participants used the computer for their study and had a computer at their disposal at home.

2.2 STIMULI

The participants were given one of four sets of eight general medical questions for which the answers could be found on the World Wide Web. The participants had to give two types of answers per question i.e., a brief answer and an extended answer. Besides, different (combinations of) modalities could be used to answer the questions. The participants had to assess for themselves which (combination of) modalities were best for a given question, and they were specifically asked to present the answers as they would prefer to find them in a QA system. To make sure they could carry out this task, they were instructed about the working of QA systems in advance. Questions and answers had to be presented in a fixed format in PowerPoint™ with areas for the question ("vraag") and the answer ("antwoord"). This programme was chosen because it has the possibility to insert pictures, film clips, and sound fragments in an answer presentation. All participants were familiar with PowerPoint™ and most of them used it on a monthly basis (51,4%).

Of the eight questions in each set, four were randomly chosen from one hundred medical questions formulated to test the IMIX QA system (e.g., how many X-chromosomes does a woman have in her body cells?). Of the remaining four questions, two were definition questions and two were procedural questions. Orthogonal to this, two questions referred explicitly or implicitly to body parts and two did not. These four question types were given to the participants in a random order. Examples of the questions were:

- Definition question + body parts: "Where is progesterone produced?" or "Where are red blood cells produced?"
- Definition question - body parts: "What are the side effects of ibuprofen?" or "What are thrombolytic drugs?"
- Procedural question + body parts: "How should a sling be applied to the left arm?" or "What should be done when having a nosebleed?"
- Procedural question - body parts: "What happens when a myelogram is taken?" or "How is a SPECT scan made?"

2.3 CODING SYSTEM

Each answer was coded as belonging to a category of the following variables: the presence of text, typographic manipulation, spatial manipulation, graphical manipulation, photos, graphics, animations, the function of these visual media[2] related to text, and the position of the visual media related to text. Our coding criteria for these variables are discussed below. To determine the reliability of the coding system, Cohen's κ (Krippendorff, 1980) was calculated. However, the Cohen's κ was not calculated for two of the variables: number of words and text. The number of words was counted automatically; therefore no agreement had to be established between the annotators. Second, text occurred in 98% of the answers. The remaining 2% of the answers were insufficient to determine the Cohen's κ. Below we will describe our criteria for coding the answers.

Number of words: The number of words was counted automatically.

Text: We distinguished the presence of textual answers (i.e., answers that contained text, possibly in combination with other media) versus non-textual answers.

[2] By visual media we mean photos, graphics, and animations.

Typographic manipulation: An answer contained typographic manipulation if the following features occurred: the use of bold, italic, underlining, or colour in the text of the answer.

Spatial manipulation: An answer contained spatial manipulation if the following features occurred: dividing the text into sections, indenting the text, using headings, or using enumeration.

Graphical manipulation: An answer contained graphical manipulation if the following features occurred: using tables, horizontal or vertical lines, arrows, or bullets.

Photos: We distinguished whether the answer contained no photo, one photo or several photos.

Graphics: We defined graphics as non-photographic, static depictions of concepts (e.g., diagrams, charts, and line drawings). We distinguished whether the answer contained no graphic, one graphic, or several graphics.

Animations: We defined animations as dynamic visuals possibly with sound (e.g., film clips and animated pictures). We distinguished whether the answer contained no animations, one animation, or several animations.

Position of visual media: We wanted to know what the position of the visual media (i.e., photos, graphics, and animations) was compared to the text within the answer presentations. We distinguished whether the visual media were in the upper, lower, left or right part of the answer area.

Function of visual media: We wanted to know what the function of the visual medium (i.e., photos, graphics, and animations) was in relation to text within the answer presentations. We distinguished three functions, loosely based on Carney & Levin (2002):

1. *Decorational function*: a visual medium has a decorational function if removing it from the answer presentation does not alter the informativity of the answer in any way. Figure 1 shows two examples of answer presentations in which the visual medium has a decorational function. The example on the left shows an answer to the question: "What are the side effects of a vaccination for diphtheria, whooping cough, tetanus, and polio?" Within the answer spatial manipulation occurs (i.e., the text is divided into sections and an enumeration is used). Also graphical manipulation occurs (i.e., bullets are used). The answer consists of a combination of text + graphic. The text describes the side effects of the vaccination while the graphic only shows a syringe. The graphic does not add any information to the answer; therefore it has a decorational function. The example on the right shows an answer to the question: "How many X-chromosomes are there in a woman's body cell?" The answer consists of a combination of text + graphic. In text the answer is given (i.e., a woman's body cell has two X-chromosomes). The answer would not be less informative if the graphic was not present.

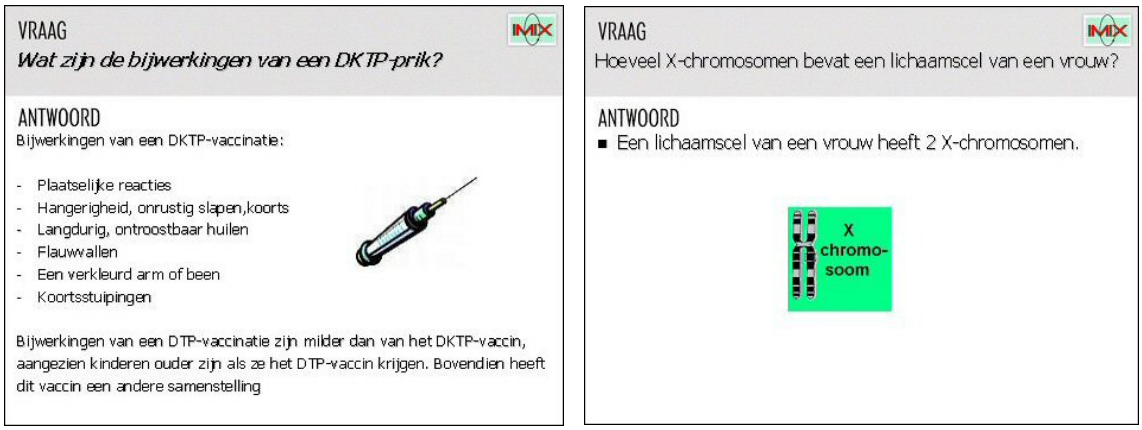

Figure 1: Examples of answer presentations with a visual medium having a decorational function

2. *Representational function*: a visual medium has a representational function if removing it from the answer presentation does not alter the informativity of the answer, but its presence clarifies the text. Figure 2 shows two examples of answer presentations in which the visual medium has a representational function. The example on the left shows an answer to the question: "What types of colitis can be distinguished?" Within the answer spatial manipulation (i.e., an enumeration is used) and graphical manipulation occurs (i.e., bullets are used). The answer consists of a combination of text + graphic. The text describes the four types of colitis and their occurrence in the intestines. This information is visualized in the graphics. The example on the right shows an answer to the question: "How should a sling be applied to the left arm?" The answer consists of three photos illustrating the procedure, which is described in more detail in the text on the right.

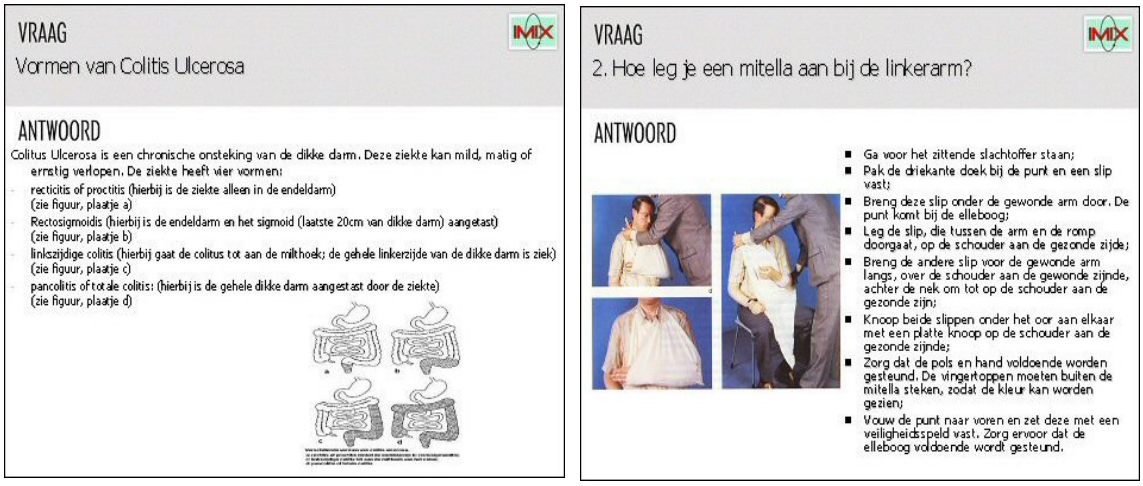

Figure 2: Examples of answer presentations with a visual medium having a representational function

3. *Additional function*: a visual medium has an additional function if removing it from the answer presentation alters the informativity of the answer. If an answer consists only of a visual medium, it automatically has an additional function. Figure 3 shows two examples of answer presentations in which the visual medium has an additional function. The example on the left shows the answer to the question: "How should a sling be applied to the left arm?" The answer consists of four graphics illustrating the procedure. The example on the right shows an answer to the question: "How can I strengthen my abdominal muscles?" The text describes some general information about abdominal exercises (i.e., an exercise program should be well balanced in which all abdominal muscles must be trained). The photos represent four exercises which can be done to strengthen the abdominal muscles.

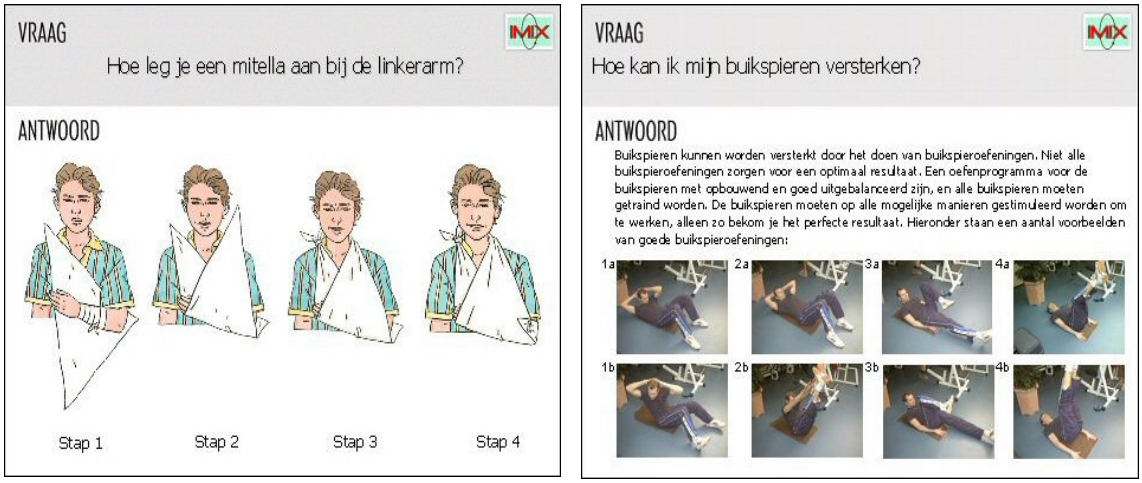

Figure 3: Examples of answer presentations with a visual medium having an additional function

2.4 ANNOTATION PROCEDURE

In total 1776 answers were collected (111 participants × 8 questions × 2 (brief, extended) answers). However, one participant gave 15 answers, resulting in one missing value. Thus, the coded corpus consisted of 1775 answers. The coding scheme (see Section 2.3) was formulated and given to six annotators (the authors plus one other annotator). The annotation was done in two steps. First, each annotator independently coded a part of the corpus to determine the adequacy of the coding scheme. Differences between the annotators were discussed, which resulted in some adjustments of the coding system. Subsequently, every annotator independently coded the same set of 112 answers. Second, every annotator independently coded a part of the total corpus (i.e., approximately 300 answers).

To compute agreement we used Cohen's κ measure. Following standard practice, Cohen's κ scores between .81 and 1.00 signify an almost perfect agreement, between .61 and .80 signify a substantial agreement, between .41 and .60 is a moderate agreement, and between .21 and .40 is a fair agreement (Rietveld & van Hout, 1993). Table I summarizes the results. The annotators corresponded in judging the occurrence of typographic manipulation. They highly corresponded in judging the occurrence of spatial manipulation, photos, graphics, and animations. Moreover, an almost perfect agreement was reached in assigning a function to the visual media, and a substantial agreement was reached in assigning a position to the visual media[3]. However, a low agreement was reached for the occurrence of graphical manipulation. A possible explanation for this result could be that the use of graphical manipulation interfered with the use of PowerPoint™. This program presents the information point by point using bullets. It was not clear whether the participants used the bullets intentionally or unintentionally to present the information. Therefore, some analysts coded the use of bullets as an occurrence of graphical manipulation and some did not, resulting in a low kappa score for this variable.

Typographic manipulation	.74
Spatial manipulation	.89
Graphical manipulation	.41
Photo's	.81
Graphics	.83
Animations	.92
Function of visual media	.83
Position of visual media	.74

Table 1: Cohen's κ scores of agreement for typographic manipulation, spatial manipulation, graphical manipulation, photos, graphics, animations, the function and the position of visual media (n = 112 answer presentations)

3 RESULTS

3.1 DESCRIPTIVE STATISTICS

Table 2 shows the frequencies of text, typographic manipulation, spatial manipulation, graphical manipulation, visual media (overall), photos, graphics, and animations in the complete corpus of coded answer presentations. Inspection of Table 2 reveals that most answer presentations contain text. Almost one in five answers contained typographic manipulation. Spatial manipulation occurred in almost half of the answer presentations and graphical manipulation occurred in one of the six answer presentations. Almost one in four answers contained one or more visual media of which graphics were most frequent and animations were least frequent. The presence of photos was between these two.

[3] The Cohen's κ for the variable "position of visual media" is based on the judgments of five annotators.

Table 3 shows the frequencies of photos and graphics related to their position.[4] The analysis of the position of visual media revealed significant effects for both photos (χ^2 (3) = 75.96, p < .001) and graphics (χ^2 (3) = 176.02, p < .001). In both cases, the medium was most often placed below the text.

Table 4 shows the frequencies of photos, graphics, and animations related to their function. Note that the answer presentations in which photos, graphics, or animations co-occurred are not shown in the table. Table 3 reveals that the distribution of photos related to their function differed significantly from chance (χ^2 (2) = 42.84, p < .001). Most photos had a representational function. Also, the distribution of graphics related to their function differed significantly from chance (χ^2 (2) = 34.50, p < .001). Most graphics had a representational function. Finally, the distribution of animations related to their function differed significantly from chance (χ^2 (2) = 63.88, p < .001). Most animations had an additional function.

Text		98.3
Typographic manipulation		18.1
Spatial manipulation		47.6
Graphical manipulation		16.7
Visual media[5]		24.0
	Photos	9.0
	Graphics	14.2
	Animations	3.6

Table 2: Frequencies of text, typographic manipulation, spatial manipulation, graphical manipulation, photos, graphics, and animations in 1775 coded answers from 111 participants (Scores are percentages of answers; n = 1775)

	Position of visual media with respect to text				
	Above text	Below text	Left of text	Right of text	Totals
Photo (n = 114)	8.8	56.1	3.5	31.6	100.0
Graphic (n =193)	3.6	61.1	2.1	33.2	100.0

Table 3: Frequencies of photos and graphics related to their position (Scores are percentages of answers)

[4] The position of animations was not taken into account because they were always added to the answer presentations with hyperlinks. Moreover, in this analysis answer presentations in which only a visual medium occurred are not taken into account.

[5] In some answers several visual media occurred (i.e., photos, graphics, and animations). These instances were counted as one occurrence of visual media. Thus the sum of the frequencies of photos, graphics, and animations does not correspond with the overall frequency of the variable visual media.

	Function of visual media			
	Decorational function	Representational function	Additional function	Totals
Photos (n = 129)	20.9	60.5	18.6	100.0
Graphics (n = 221)	15.4	46.6	38.0	100.0
Animations (n = 48)	2.1	10.4	87.5	100.0

Table 4: Frequencies of photos, graphics, and animations related to their function (Scores are percentages of answers)

3.2 BRIEF AND EXTENDED ANSWERS

As expected the type of answer (brief vs. extended) affected the answer presentation. The type of answer had a significant effect on the mean number of words used ($t(1726) = 30.39$, $p < .001$). The mean number of words used in brief answers was 24 words (Std = 23,02), while the mean number of words used in extended answers was 106 words (Std = 76,20). Table 5 shows the frequencies and χ^2 statistics of the presence of text, typographic manipulation, spatial manipulation, graphical manipulation, visual media (overall), photos, graphics, and animations within the brief and extended answers. The results showed that there was a significant difference in the presence of all the variables within the answer type. Inspection of Table 5 reveals that they all occurred more frequently within the extended answers.

Table 6 shows the frequencies and χ^2 statistics of the functions of visual media related to brief and extended answers. The results showed that the overall distribution of the functions of visual media within the answer type differed significantly ($\chi^2(2) = 31.47$, $p < .001$). Visual media with a decorational function occurred significantly more often in brief answers than in extended answers. Visual media having a representational function occurred significantly more often in extended answers. Finally, visual media having an additional function occurred significantly more often in brief answers.

	Brief answers (n = 888)	Extended answers (n = 887)	χ^2 statistics
Text	97.5	99.0	$\chi^2(1) = 5.53$, $p < .025$
Typographic manipulation	9.8	26.5	$\chi^2(1) = 83.30$, $p < .001$
Spatial manipulation	23.9	71.4	$\chi^2(1) = 401.24$, $p < .001$
Graphical manipulation	8.9	24.5	$\chi^2(1) = 77.40$, $p < .001$
Visual media	11.0	38.0	$\chi^2(1) = 174.30$, $p < .001$
Photos	4.8	13.1	$\chi^2(1) = 36.90$, $p < .001$
Graphics	5.5	22.9	$\chi^2(1) = 109.98$, $p < .001$
Animations	.9	6.3	$\chi^2(1) = 37.40$, $p < .001$

Table 5: Frequencies and χ^2 statistics of the presence of text, typographic manipulation, spatial manipulation, graphical manipulation, visual media (overall), photos, graphics, and animations related to the brief and extended answers (Scores are percentages of answers; n = 1775)

	Brief answers (n = 98)	Extended answers (n = 338)	χ^2 statistics
Decorational function	25.5	13.3	χ^2 (1) = 5.71, p < .025
Representational function	21.4	53.3	χ^2 (1) = 125.78, p < .001
Additional function	53.1	33.4	χ^2 (1) = 22.55, p < .001
Totals	100.0	100.0	

Table 6: Frequencies of the function of visual media related to brief and extended answers (Scores are percentages of answers; n = 436)

3.3 DEFINITION AND PROCEDURAL QUESTIONS WITH AND WITHOUT REFERENCE TO BODY PARTS

We were interested whether different types of questions were related to different answer presentations. Therefore we analyzed a subset of the medical questions (i.e., the definition and procedural questions with and without reference to body parts). The results indicated that the type of question had a significant effect on the mean number of words used in the answer presentation (F (3, 484.63) = 9.28, p < .001)[6]. Post hoc tests indicated that the answers of procedural questions consisted of more words than the answers of definition questions irrespective of reference to body parts.

Table 7 shows the frequencies and χ^2 statistics of the presence of text, typographic manipulation, spatial manipulation, graphical manipulation, visual media (overall), photos, graphics, and animations within the definition and procedural questions and within questions with and without reference to body parts. The results showed that the distribution of typographic manipulation within the question types did not differ: typographic manipulation occurred equally within all question types. However, the distribution of all other variables within the question types differed significantly. Text occurred most frequently within definition questions with reference to body parts and procedural questions without reference to body parts. Spatial and graphical manipulation were most frequent in procedural questions with reference to body parts. The use of visual media was also most frequent within this type of questions. Finally, photos and animations occurred more often in answers to procedural questions with reference to body parts. However, graphics occurred more often in answers to *definition* questions with reference to body parts.

Table 8 shows the frequencies and χ^2 statistics of the functions of visual media within definition and procedural questions and within questions with and without reference to body parts. The results show that the functions of visual media differed significantly within the question types (χ^2 (6) = 91.84, p < .001). Table 8 shows that visual media with a decorational function occurred most often in definition questions *without* reference to body parts, and that visual media with a representational function occurred most often in definition questions *with* reference to body parts. Finally, visual media having an additional function occurred most often in *procedural* questions with reference to body parts.

4 CONCLUSION

In this paper we described a production experiment following a cognitive engineering approach in order to gain knowledge on which modality combinations are used in manually created answers. A total of 1775 answers to different medical questions were collected. These answers were coded as belonging to a category of the following variables: text, typographic manipulation, spatial manipulation, graphical manipulation, photos, graphics, animations, the function and the position of the visual media related to text. To determine the reliability of this coding scheme, six annotators coded part of the data. The results of this reliability analysis indicated that for most variables the annotators corresponded highly in their judgments.

[6] Tests for significance were performed using a Welch's analysis of variance (ANOVA), as the variances were not equal. A significance threshold of .05 was used and for post hoc tests the Tukey HSD method was used.

	Definition questions (n = 443)		Procedural questions (n = 444)		
	Body parts (n = 222)	¬ Body parts (n = 221)	Body parts (n = 222)	¬ Body parts (n = 222)	χ^2 statistics
Text	99.5	99.1	94.1	99.5	χ^2 (3) = 24.61, p < .001
Typographic manipulation	22.5	16.3	18.0	17.1	χ^2 (3) = 3.421, p = .33
Spatial manipulation	38.7	53.4	61.3	41.4	χ^2 (3) = 29.47, p < .001
Graphical manipulation	9.5	20.8	29.7	9.5	χ^2 (3) = 44.83, p < .001
Visual Media	31.1	10.4	46.8	32.4	χ^2 (3) = 70.84, p < .001
Photos	4.5	5.9	24.3	19.8	χ^2 (3) = 55.73, p < .001
Graphics	28.4	5.0	13.1	11.7	χ^2 (3) = 52.29, p < .001
Animations	.5	.9	13.5	5.0	χ^2 (3) = 51.74, p < .001

Table 7: Frequencies and χ^2 statistics of the presence of text, typographic manipulation, spatial manipulation, graphical manipulation, visual media (overall), photos, graphics, and animations within the definition and procedural questions and within questions with and without reference to body parts.

	Definition questions (n = 92)		Procedural questions (n = 177)		
	Body parts (n = 69)	¬ Body parts (n = 23)	Body parts (n = 105)	¬ Body parts (n = 72)	χ^2 statistics
Decorational function	5.8	65.2	4.8	5.6	χ^2 (3) = 12.29, p < .01
Representational function	63.8	21.8	41.0	54.2	χ^2 (3) = 31.78, p < .001
Additional function	30.4	13.0	54.2	40.2	χ^2 (3) = 55.09, p < .001
Totals	100.0	100.0	100.0	100.0	

Table 8: Frequencies and χ^2 statistics of the functions of visual media related to the definition and procedural questions and to questions with and without reference to body parts (Scores are percentages of answers; n = 269)

A first analysis of the data showed that the participants used combinations of text and visual media to present their answers. Almost one in four answers contained one or more visual media. Moreover, significant differences were found in the distribution of photos, graphics, and animations related to their function. Photos often had a representational function: they visually represented the information mentioned in the text. Animations often had an additional function because they present the information dynamically as opposed to photos. Graphics often had either a representational or an additional function. A possible explanation for this result could be that graphics are more diverse. While some graphics visually represent the information mentioned in text, other graphics represent information in such a way (e.g., the presence of arrows or charts) that they contain more information than mentioned in the text.

As expected the type of answer (brief vs. extended) affected the answer presentation. Extended answers consisted of more words than the brief answers, but also word manipulation, spatial manipulation

and graphical manipulation were more frequent in the extended answers. A possible explanation for this result could be that presenting more text affects the readability. Typographic manipulation, spatial manipulation, and graphical manipulation could help to make the text more transparent and thus more readable. Also visual media were more frequent in the extended answers. Within brief answers, most frequent were visual media with a decorational and additional function whereas visual media with a representational function were more frequent within extended answers. A possible explanation for this result could be that when the answer does not contain much text, it is likely that the visual medium will have an additional function (i.e., it expresses more information). When the answer contains much text, it is likely that the visual medium will have a representational function (i.e., it represents the information mentioned visually).

The type of question also affected the answer presentation. Answers to procedural questions consisted of more words. Besides, spatial and graphical manipulation occurred more frequently in answers to this type of questions. A possible explanation for this result could be that procedural information consists of several steps that have to be described. Moreover, dividing the text into sections or using headings may help the user to see when one step ends and another begins (Ganier, 2004). The distribution of visual media differed significantly within the question types. Photos and animations occurred most often in answers to procedural questions with reference to body parts. These visual media may help to visualise the steps of a procedure. However, graphics occurred most often in answers to definition questions with reference to body parts. As mentioned earlier, graphics are more diverse making them perhaps more suitable for other question types. For example, the definition question "Where is testosterone produced?" may be more clearly visualized with a graphic in which different parts of the male reproductive system are illustrated.

The first results of this production experiment following the cognitive engineering approach showed that users do make use of multiple media in their information presentations and that the design of these presentations is affected by the answer and question type. However what is not clear is which kind of multimodal information presentation users would prefer. This will be tested in a second experiment in which users will rate the answer presentations collected in this production experiment. Based on the results of this experiment, we intend to develop a set of design principles for multimodal answer presentation in the IMIX medical QA system.

ACKNOWLEDGEMENTS

The current research is performed at the Communication and Cognition group at Tilburg University and the Human Media Interaction group of the University of Twente within the IMOGEN (Interactive Multimodal Output GENeration) project. IMOGEN is a project within the Netherlands Organisation for Scientific Research (NWO) research programme on Interactive Multimodal Information eXtraction (IMIX). The authors would like to thank Paul Flapper for his help in setting up the experiment, Jurry de Vos for his help in annotating the data, and Jeroen Geertzen for his help with the statistics.

REFERENCES

André, E. (2000). The generation of multimedia presentations. In R. Dale, H. Moisl, & H. Somers (Eds.) *A Handbook of Natural Language Processing: Techniques and Applications for the Processing of Language as Text*. Marcel Dekker Inc., NY. pp. 305-327.

Arens, Y., Hovy, E. & Vossers, M. (1993). On the knowledge underlying multimedia presentations. In M. T. Maybury (Ed.) *Intelligent Multimedia Interfaces*, AAAI Press, Menlo Park, CA, pp. 280 - 306.

Bernsen, N. (1994). Foundations of multimodal representations. A taxonomy of representational modalities. *Interacting with Computers 6* (4), pp. 347-371.

Bosma, W. (2005). Extending answers using discourse structure. In *Proceedings of the RANLP Workshop on Crossing Barriers in Text Summarization Research*. H. Saggion & J.-L. Minel (Eds), Incoma Ltd., Borovets, Bulgaria, pp. 2-9.

Carney, R. & Levin, J. (2002). Pictorial illustrations still improve students' learning from text. *Educational Psychology Review 14*(1), pp. 5-26

Ganier, F. (2004). Factors affecting the processing of procedural instructions: implications for document design. *IEEE Transactions on Professional Communication*. *47* (1), pp. 15 – 26.

Heiser, J., Phan, D., Agrawala, D., Tversky, B. & Hanrahan, P. (2004). Identification and validation of cognitive design principles for automated generation of assembly instructions. In *Proceedings of the Working Conference on Advance Visual Interfaces*, ACM Press, NY, pp. 311-319.

Heller, R., Martin, C., Haneef, N. & Gievka-Krliu, S. (2001). Using a theoretical multimedia taxonomy framework. *ACM Journal of Educational Resources in Computing 1* (1), pp. 1-22.

Hooijdonk, C.M.J. van & Krahmer. E. (submitted). Information modalities for procedural instructions: the influence of text, static and dynamic visuals on learning and executing RSI exercises.

Krippendorff, K. (1980). *Content Analysis: An Introduction to its Methodology*. Sage Publications, Beverly Hills, CA, USA.

Marcus, N., Cooper, M., & Sweller, J. (1996). Understanding instructions. *Journal of Educational Psychology, 88* (1), pp. 49-62.

Maybury, M. & Lee, J. (2000). Multimedia and multimodal interaction structure. In M. Taylor, F. Néel & D. Bouwhuis (Eds.). *The Structure of Multimodal Dialogue II*, John Benjamins, Amsterdam, pp. 295-308.

Mayer, R. (1989). Systematic thinking fostered by illustrations in scientific text. *Journal of Educational Psychology 81* (2), pp. 240-246

Mayer, R. & Gallini,. J. (1990). When is an illustration worth a thousand words? *Journal of Educational Psychology, 82,* pp. 715-726,

Mayer, R. & Moreno, R. (2002). Aids to computer-based multimedia learning. *Learning and Instruction, 12*, 107-119.

Mayer, R. E. (2005). *The Cambridge Handbook of Multimedia Learning*. Cambridge [etc.]: Cambridge University Press.

Michas, I. & Berry, D. (2000). Learning a procedural task: effectiveness of multimedia presentations, *Applied Cognitive Psychology, 14*, pp. 555-575.

Oviatt, S., Coulston, R., Tomko, S., Xiao, B., Lunsford, R., Wesson, M. & Carmichael, L. (2003). Toward a theory of organized multimodal integration patterns during human-computer interaction. In *Proceedings of the 5th International Conference on Multimodal Interfaces*, Vancouver, British Columbia, Canada, pp. 44–51.

Rietveld, T. & Hout, R. van (1993). *Statistical Techniques for the Study of Language and Language Behavior*. Berlin: Mouton de Gruyter.

Sutcliffe, A. (1997). Task-related information analysis. *Int. Journal of Human Computer Studies, 47* (2), 223-257.

Schwan, S. & Riempp, R. (2004). The cognitive benefits of interactive videos: learning to tie nautical knots. *Learning and Instruction, 14*, pp. 293-305.

Theune, M., Schooten, B. van, Akker, R. op den, Bosma, W., Hofs, D., Nijholt, A., Krahmer, E., Hooijdonk, C. van & Marsi, E. (to appear). Questions, pictures, answers: introducing pictures in question-answering systems. To appear in *Proceedings of the Tenth International Symposium on Social Communication*, 22-26 January 2007, Santiago de Cuba, Cuba.

Tversky, B., Morrison, J. & Betrancourt, M. (2002). Animation: can it facilitate? *Int. Journal of Human Computer Studies, 57*, 247-262.